DOI:
10.1007/s10514-019-09897-6

Document Version
Peer reviewed version

Joint Optimization Based on Direct Sparse Stereo Visual-Inertial Odometry

Shuhuan Wen[a,b], Yanfang Zhao[a], Hong Zhang (Fellow, IEEE)[b,*], Hak Keung Lam (Senior Member, IEEE)[c], Luigi Manfredi[d]

[a]*Key Lab of Industrial Computer Control Engineering of Hebei Province, Yanshan University, Qinhuangdao, China*
[b]*Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8, Canada*
[c]*Department of Informatics, King's College London, 30 Aldwych, London, WC2B 4BG, United Kingdom*
[d]*Institute for Medical Science and Technology (IMSaT), University of Dundee, United Kingdom*

Abstract

This paper proposes a novel fusion of an inertial measurement unit (IMU) and stereo camera method based on direct sparse odometry (DSO) and stereo DSO. It jointly optimizes all model parameters within a sliding window, including the inverse depth of all selected pixels and the internal or external camera parameters of all keyframes. The vision part uses a photometric error function that optimizes 3D geometry and camera pose in a combined energy functional. The proposed algorithm uses image blocks to extract neighboring image features and directly forms measurement residuals in the image intensity space. A fixed-baseline stereo camera solves scale drift. IMU information is accumulated between several frames using manifold pre-integration and is inserted into the optimization as additional constraints between keyframes. The scale and gravity inserted are incorporated into the stereo visual inertial odometry model and are optimized together with other variables such as poses. The experimental results show that the tracking accuracy and robustness of the proposed method are superior to those of the state-of-the-art fused IMU method. In addition, compared with previous semi-dense direct methods, the proposed method displays a higher reconstruction density and scene recovery.

Keywords: Direct sparse odometry, IMU pre-integration, sliding window optimization, 3D reconstruction

1. Introduction

Recently, simultaneous localization and mapping (SLAM) has been a popular research topic in robotics, because it is a fundamental building block for many emerging technologies such as self-driving cars [1], robotic navigation [2], unmanned aerial vehicles (UAVs), virtual reality (VR), and augmented

*Corresponding author
Email addresses: swen@ysu.edu.cn (Shuhuan Wen), FangYZhao@163.com (Yanfang Zhao), hzhang@ualberta.ca (Hong Zhang (Fellow, IEEE)), hak-keung.lam@kcl.ac.uk (Hak Keung Lam (Senior Member, IEEE)), l.manfredi@dundee.ac.uk (Luigi Manfredi)

reality (AR). Pose tracking has attracted significant attention in computer vision. While traditional robotic systems such as self-driving cars have largely relied on LiDAR to actively sense the environment and perform self-localization and mapping, visual SLAM and odometry algorithms have greatly improved in terms of performance. Compared with other sensors, the camera and IMUs are inexpensive, ubiquitous and complementary, and can be combined to work jointly. Visual sensors provide rich information for robust visual tracking. Using vision, the full 3D rotation and translation of a robot can be observed in an environment with sufficient features. IMUs can measure accelerations and angular velocities at high frame-rates in order to maintain the tracking of feature points in such cases as when the camera points at a wall with poor texture. IMUs can also observe acceleration resulting from gravity and extract an absolute horizontal reference in the environment.

There are direct and indirect methods for solving SLAM and visual inertial odometry (VIO). Indirect methods (i.e, feature-based methods) operate in two steps. First, a set of feature observations are extracted from the image. Second, the camera position and scene geometry are estimated in a probabilistic model. The two steps are usually independent. In the second step, the re-projection error can be used to remove outlier points in data association and to correct the matching result (such as the EM-like method in [3]). A direct method uses actual sensor photometric values received from the gradient direction over a period of time as a measurement of a probabilistic model and regards data association and pose estimation as a unified nonlinear optimization problem. Furthermore, an indirect method calculates the re-projection error of the camera pose and the positions of feature points in the robot the map, while a direct method calculates the photometric error. The so-called photometric error means that the minimized objective function is usually determined by the error between the images rather than the geometric error after re-projection. In this paper, we propose a tightly coupled direct method for VIO. Motion estimation and 3D reconstruction are important technologies for robots. Compared with other sensors, the camera and IMU are inexpensive and lightweight. Owing to these advantages, the camera and IMU have received wide attention. However, when faced with low-texture areas or during fast maneuvering, the current visual SLAM and VO methods lack robustness. An IMU can improve this robustness by providing accurate short-term motion constraints.

In this paper, we propose a novel stereo VIO method. It is based on DSO [4], and uses bundle adjustment (BA) to minimize photometric error and optimize 3D geometry and camera poses in a combined energy functional. Camera pose, velocity and IMU biases are simultaneously estimated by minimizing a cost function which combines visual energy functional with inertial energy functional. At the same time, the external parameter which is the rotation and translation between camera and IMU can also realize real-time online correction. This method is particularly beneficial for direct methods

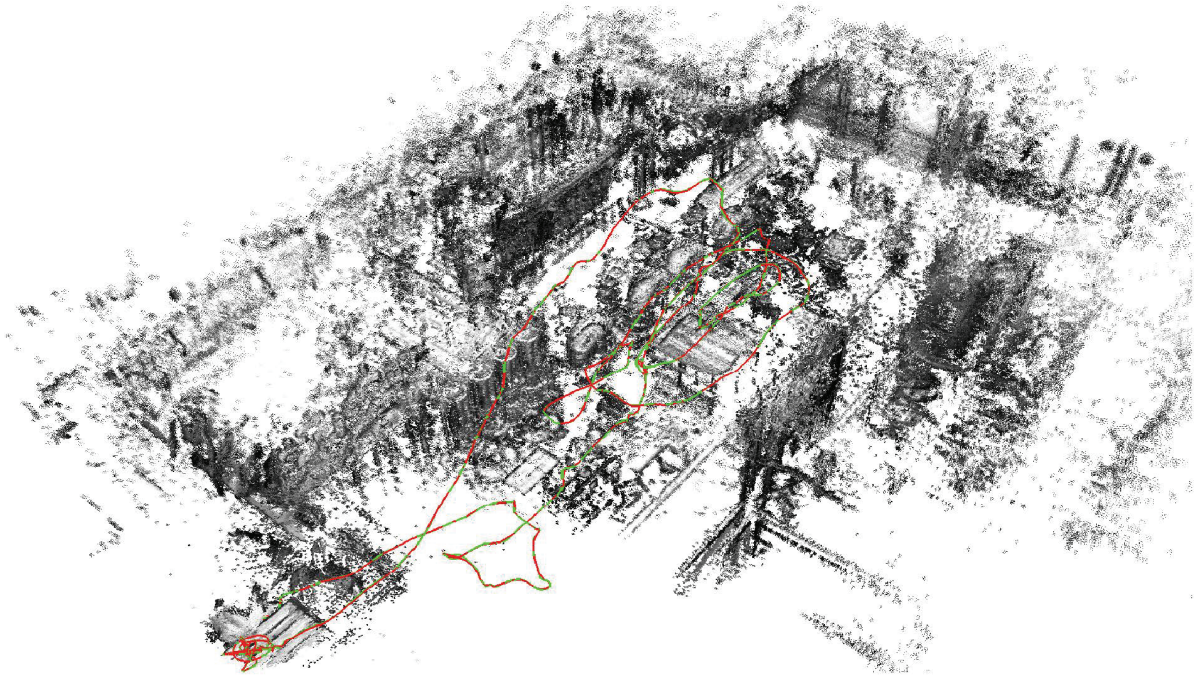


Figure 1: Recostruction of EuRoC-dataset tighted fusion of IMU measurement with direct image alignment.

because the error function is highly non-convex and good initialization is essential. Compared with monocular visual odometry, stereo visual odometry can compute pixel depths using triangulation. An IMU fused with vision also enables us to observe the scale of the environment. Fig. 1 shows the output of our stereo VIO algorithm when running in an indoor dataset.

The rest of the paper is structured as follows. In Sect. 2, we discuss the relevant literature. The motivation and system overview are discussed in Sect. 3. Implementation details and experimental evaluations are presented in Sect. 4. Finally, the paper is concluded in Sect. 5.

2. Related work

Recently, motion estimation combining cameras and IMUs has been a popular research topic. In this section, we provide an overview of the vision-only and VIO methods for motion estimation, and we also discuss the direct vs. feature-based approach.

The first work on direct visual-odometry was reported in 2007 [5]. Since then, direct methods have been used in the RGB-D camera [6], as they directly provide the required pixel-wise depth as sensor measurement. Recently, direct methods have also become popular in monocular cameras. Newcombe et al. proposed the dense tracking and mapping method called DTAM [7], Forster et al. proposed fast

semi-direct monocular visual odometry methods called SVO [8], Engle et al. proposed LSD-SLAM [9] and DSO [4]. Direct methods can compute geometry and motion directly from the images and thus they do not involve the intermediate step of keypoint detection and matching. Indirect methods such as ORB-SLAM [10][11] are based on detected keypoints. Keyframe-based methods such as PTAM [12] perform motion and 3D structure estimation in parallel.

Because vision and IMU sensors can compensate each other's weaknesses, there have been research studies on fusing vision with an IMU. IMUs can overcome the limitation of vision-based systems, provide valuable short-term motion constraints, and make roll, pitch and scale of robot pose observable. In early work, vision was treated as an independent 6 DOF sensor and fused with IMUs measurement in a filter framework [13]. There are loosely coupled and tightly coupled methods for the fusion of the IMU and vision. If a loosely coupled [14] method using an existing vision-only algorithm is not modified, it can be easily replaced by another method. However, it cannot benefit from the availability of IMU data to work with for vision.

Thus some recent works used a tightly coupled method, which regards VIO as one integrated estimation problem, two sensors are used in the optimization-based back-end. For example, Leutenegger et al. proposed a keyframe-based VIO [15], Qin et al. proposed VINS-Mono [16], a robust and versatile monocular visual-inertial state estimator, and Mur-Artal et al. proposed a tightly coupled visual inertial simultaneous localization and mapped system [17] that is able to close loops and reuse its map to achieve zero-drift localization in the mapping areas.

However, visual-only algorithms and visual inertial algorithms have attracted the attention of many researchers. The accuracy and robustness of DSO [4] outperformed state-of-the-art monocular SLAM algorithms such as ORB-SLAM [10] on a reasonably large dataset for monocular camera tracking [18]. The direct sparse VIO method using dynamic marginalization [19] proposed by Stumberg outperformed the VI ORB-SLAM [20]. In [19], the stereo method [2] was better than monocular VIO. In [21], Wang et al. proposed a stereo DSO method. There was a higher reconstruction density than in feature-based methods, and the vision-only algorithm was not sensitive to fast motion, while IMU measurements could overcome this problem.

An energy-based method was proposed in [22]. This method combines IMU measurements with direct tracking of a parse subset of points in the image. Energy-based methods jointly optimize camera and IMU parameters by a combine energy function. They can obtain more effective data which can make the system more robust. In this paper, we combine stereo DSO methods and IMU pre-integration in an energy function, and jointly optimize the parameters of the energy function.

3. Direct sparse stereo visual inertial odometry

Stereo VIO is a system based on the iterative minimization of photometric errors and IMU measurement errors, such as in [23], the fused IMU method has several advantages. The proposed method is on the basis of [19] and [21]. We use the nonlinear optimization method in the monocular_VIDSO [19]. However we estimate depth by stereo alignment and provide the priori for IMU initialization, which is different from [19]. We use the fixed baseline [21] for the front-end odometry of our method. We combine the photomatric error with IMU measurement errors to form a single optimization function, which is different from [21]. An overview of the proposed system in this paper is shown in Fig 2. The system starts with sensor measurements, in which direct coarse tracking and IMU measurements between two contiguous frames are pre-integrated.

To improve VIO based on nonlinear optimization, the initialization process provides all crucial values, including pose, velocity, gravity, vector, gyroscope bias, and 3D feature location. VIO with localization models tightly fuses IMU pre-integration measurements, feature observations, and re-detected features. Finally, the optimization module of the pose graph adopts the localization results as verified geometrically and performs global optimization to eliminate drift. Every module has different running rates and can perform real-time and reliable operation at all times. Compared to [21], we use global BA-based optimization to replace structure-less vision error items. The proposed method estimates robot pose and scene depths by minimizing the energy function

$$E = \lambda E_{stereo} + E_{IMU} \tag{1}$$

where E_{stereo} is the stereo photometric error and E_{IMU} is the IMU measurements error.

The direct sparse stereo VIO system includes three main parts:

1. Coarse tracking is performed for each frame, and the nearest frame is estimated by combining the direct image alignment with the IMU measurement error.
2. An optimal visual inertial BA is used to estimate the geometry and poses of all active keyframes when a new keyframe is created.
3. A sliding window is created, and old keyframes and 3D points are marginalized out using a Schur complement.

3.1. Notation

In this paper we will use the following notations: light lower case λ and bold lower case letters denote vectors ($\mathbf{t}$) and scalar ($\mathbf{u}$), and bold upper case letters denote matrices ($\mathbf{R}$). Upper case letters are used to represent functions (I).

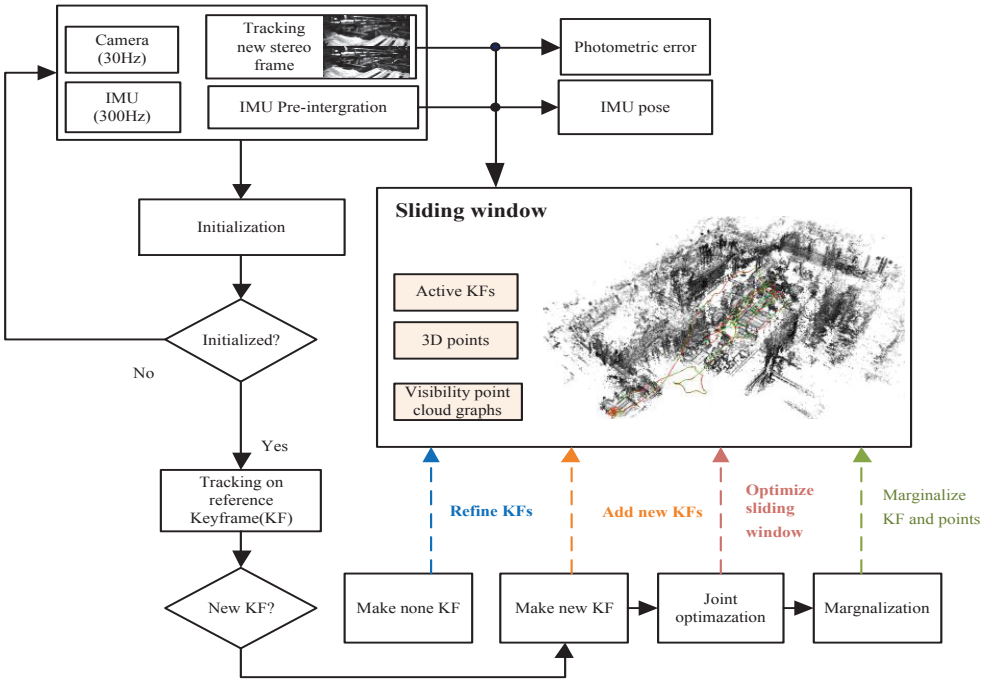


Figure 2: System overview.

Camera calibration metrics are expressed as $\mathbf{K}$. Camera poses are expressed by matrices of the special Euclidean group $\mathbf{T}_i \in SE(3)$, which transform 3D coordinates from the camera coordinate system to the world coordinate system. We denote Lie algebra elements as $\xi \in \mathfrak{se}(3)$, and use them to apply small increments to the 6D pose $\xi'_{i-j} = \xi_{i-j} \boxplus \xi := \log(\exp(\xi_{i-j}) \exp(\xi))$, where $\xi \in \mathbb{R}^6$. $\prod_{\mathbf{K}}$ and $\prod_{\mathbf{K}}^{-1}$ are used to denote camera projection and back-projection functions, respectively. In this paper, a 3D point is represented by its image coordinate $\mathbf{p}$ and inverse depth d_p relative to its host keyframe. The keyframe is the frame of the selected point—pixels with distinct gradient descent. The inverse depth parameterization is good when the errors of the images are Gaussian distributions. We denote the world as a fixed-inertial-coordinate frame with gravity acting in the negative Z direction axis. We also assume that the transformation from the camera to IMU frame $\mathbf{T}_{IMU-cam}$ is fixed and calibrated in advance.

3.2. Stereo direct sparse odometry

3.2.1. Direct image alignment

A point set, $\mathcal{P}_i$, in reference frame I_i, is assumed to be observed in another frame I_j. The direct image alignment can be formulated as

$$E_{ij} = \sum_{p \in \mathcal{P}_i} \omega_p ||I_j [\mathbf{p}'] - I_i [\mathbf{p}] ||_\gamma, \tag{2}$$

where $||\cdot||_\gamma$ is the Huber norm and ω_p is a weight that is inversely proportional to image gradient magnitude

$$\omega_p = \frac{c^2}{c^2 + ||\nabla I_i(\mathbf{p})||_2^2}, \tag{3}$$

where c is a constant, and $\mathbf{p}'$ is the projection of $\mathbf{p}$ in I_j, which is calculated by

$$\mathbf{p}' = \prod_{\mathbf{K}}(T_{ji} \prod_{\mathbf{K}}^{-1}(\mathbf{p}, d_p)), \tag{4}$$

where d_p is the inverse depth of $\mathbf{p}$. The expression that transforms a point from frame i to frame j is

$$T_{ji} = \begin{bmatrix} R_{ji} & t \\ 0 & 1 \end{bmatrix} = T_j^{-1} T_i. \tag{5}$$

General direct methods tend to use as many pixels from each image as possible. Although this is computationally intensive, the system can converge quickly. Therefore, in [4], a strategy was proposed to select a fixed number of points from each frame and evenly across all regions with a sufficient gradient. The neighborhood of each selected point is used to calculate the photometric error in Eq. (2). In this paper, we adopt the same method, but we use the stereo image pair to verify the selected points and perform a depth initialization similar to [21].

Because the photometric error is calculated directly on the pixel intensities, it is sensitive to sudden illumination changes between consecutive frames. In the ideal case, as well as the camera response time of each frame, the camera response function is directly accessible from the hardware [18], which can be used to correct the results. If this information is not available, two parameters a_i and b_i for each image are used to model the affine brightness change [4]. The energy function in Eq. (2) is then modified to

$$E_{ij} = \sum_{p \in \mathcal{P}_i} \sum_{\tilde{p} \in \mathcal{N}_p} \omega_{\tilde{p}} ||(I_j[\tilde{\mathbf{p}}'] - b_j) - \frac{e^{a_j}}{e^{a_i}}(I_i[\tilde{\mathbf{p}}] - b_i)||_\gamma, \tag{6}$$

where $\mathcal{N}_p$ is the eight-point pattern of $\mathbf{p}$ in [4] and $\tilde{\mathbf{p}}'$ is the projection of the pattern point $\tilde{\mathbf{p}}$ into I_j, a_i, b_i, a_j and b_j are estimated in the windowed optimization.

The global photometric error contend of all points and frames is

$$E_{stereo} = \sum_{i \in \mathcal{F}} \sum_{p \in \mathcal{P}_i} \sum_{j \in obs(\mathbf{p})} E_{ij}, \tag{7}$$

where $\mathcal{F}$ is the set of keyframes in the current windows, $\mathcal{P}_i$ is the sparse set of points in keyframe i, and $obs(p)$ is the set of keyframes in $\mathcal{F}$ that can observe $\mathbf{p}$.

3.2.2. Frame management

The proposed method keeps an active window of $\mathcal{N}_f$ keyframes ($\mathcal{N}_f = 7$ in our experiments). Every new frame is initially tracked for these reference frames (see Step 1 below). Then, it is either discarded

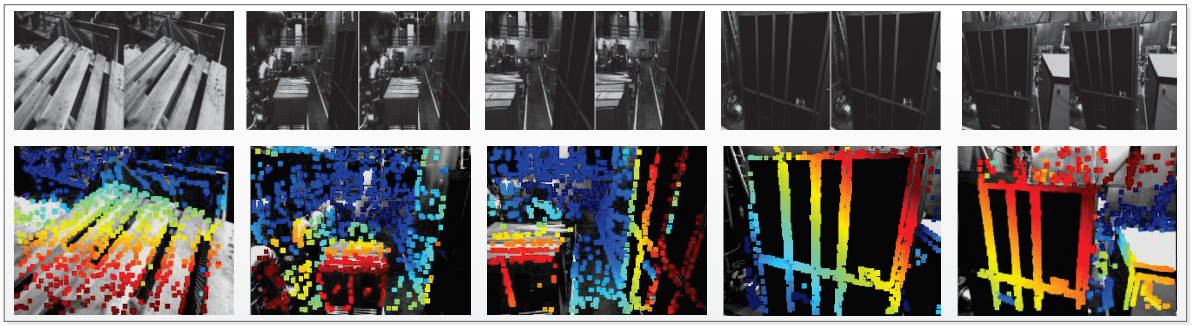


Figure 3: Examples of depth maps used for initial frame tracking (EuRoC dataset). The top row is the original images, the bottoms row is the color-coded depth maps.

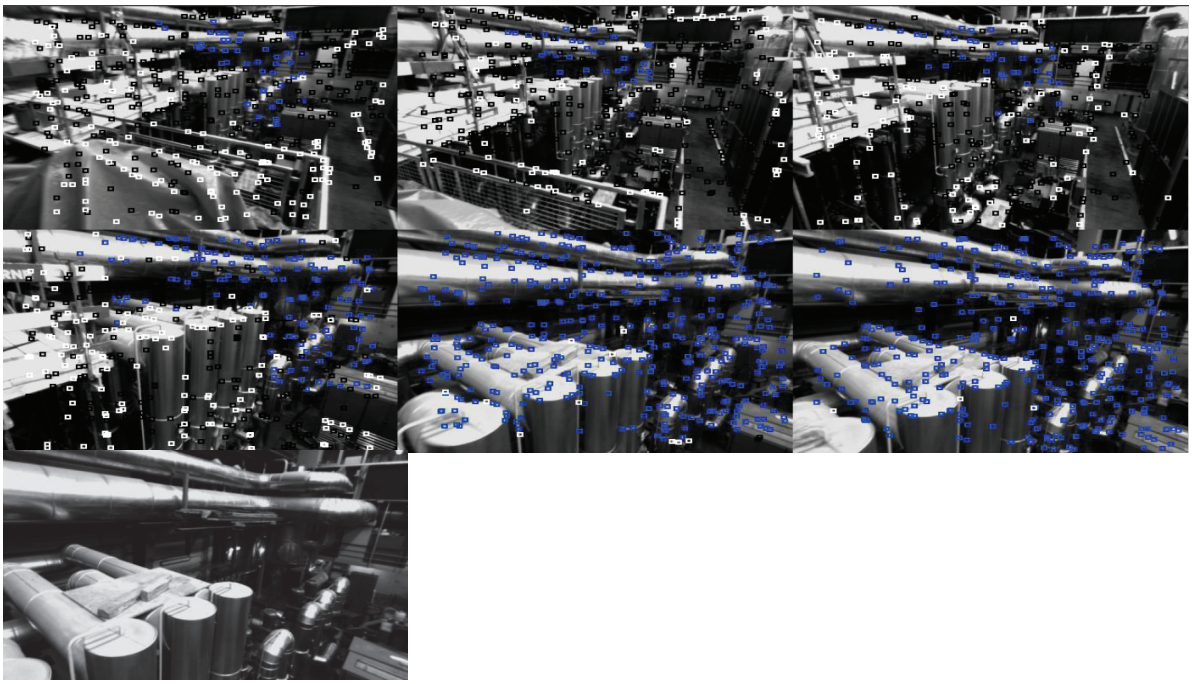
or used to create a new keyframe. Once a new keyframe and the corresponding new points are created, the total photometric error in Eq. (7) is optimized. Then, we marginalize one or more frames (see Step 2 below).

Step 1.Coarse tracking and initialization. Whenever a new stereo frame is input to the system, direct image alignment is used to track the latest keyframe in the sliding window. By using traditional two-frame direct image alignment, we use a multi-scale image pyramid and a constant motion model to initialize and track new frames. Fig. 3 shows examples of depth maps. A constant-motion model is used to obtain the initial pose of a new frame, and all observable points in the sliding window are projected into the new frame. By minimizing the visual energy function in Eq. (6), the poses of all new frames are optimized, while the value of depth is kept fixed.

In previous work [9][4][24], direct monocular visual odometry usually needed a certain pattern of initial camera movement, and then the entire system could be initialized. In this paper, we use static stereo matching to estimate a semi-dense depth map for the first frame. At this stage, the transfer factor of the affine luminance between stereo image pairs is unknown, and the corresponding relationship is searched along the horizontal epipolar line. This step also provides a prior for the initialization of the IMU.

In order to select evenly distributed points on the image and only select points with a sufficient image gradient, the image is divided into small blocks, and an adaptive threshold is calculated for each block. If the selected point exceeds the threshold of a block, the point that has the largest absolute gradient in its neighborhood is selected. Then, we obtain the inverse depth values of the point in the first frame and tracked the second frame according to Eq. (6). Furthermore, we use the size of the blocks that is proportional to the size of the image to improve the observability of the image. During the

Figure 4: Keyframe marginalization.



The black points are the marginalized points, the white points represent the candidated points, and the blue points are the host points.

initialization, we use static stereo matching with normalized cross-correlation (NCC) to obtain depth between frames, which can increase the tracking accuracy.

Step 2.Keyframe creation and marginalization. When a new stereo frame is successfully tracked, more key frames are obtained. Then the redundant keyframes are removed by marginalizing, and finally, useful keyframes are obtained. There are three rules for determining keyframes:

1. Use the mean squared optical flow $f := (\frac{1}{n}\sum_{i=1}^{n}||\mathbf{p}-\mathbf{p}'||^2)^{\frac{1}{2}}$ to judge the change in the angle of view.
2. The occlusion is judged by the mean optical flow $f_t := (\frac{1}{n}\sum_{i=1}^{n}||\mathbf{p}-\mathbf{p}'_t||^2)^{\frac{1}{2}}$ when the camera does not rotate, where $\mathbf{p}'_t$ is the warped point position $\mathbf{R}=\mathbf{I}_{3\times 3}$.
3. The relative brightness factor between two frames $a := |\log(e^{a_j-a_i}t_j t_i^{-1})|$ is used to judge the change in exposure time of the camera.

Now, the three vectors are obtained in initial alignment. If $w_f f + w_{f_t} f_t + w_a a > 1$, a new keyframe is selected. w_f, w_{f_t} and w_a are the weighted versions of the three vectors.

When the old points are removed from the active window by marginalization, the candidate points are activated and added to the joint optimization. Each activated point is hosted in one keyframe and is observed by several other keyframes in the active window. Every time an active point is observed in another keyframe, so it creates a photometric energy factor, defined as the inner part of Eq. (6)

$$E'_{ij} = \omega_{\tilde{p}} ||(I_j [\mathbf{p}'] - b_j) - \frac{e^{a_j}}{e^{a_i}}(I_i [\mathbf{p}] - b_i)||_\gamma. \tag{8}$$

Then Eq. (7) is written as

$$E_{stereo} = \sum_{i \in \mathcal{F}} \sum_{p \in \mathcal{P}_i} \sum_{j \in obs(\mathbf{p})} E'_{ij}. \tag{9}$$

Fig. 4 shows an example of a scene that highlights the active set of points and frames. In Fig. 4, there are six old keyframes in the optimization window. The black points are the marginalized points, the white points represent the candidate points, and the blue points are the host active points.

3.3. IMU integration

In a vision system, IMU data can be used to observe the metric scale and gravity direction. Thus, the parameters in our visual inertial system, including scale, gravity, velocity and gyroscope bias, are jointly optimized together with the other values such as poses and scenes geometry. Next, we introduce a nonlinear dynamic model. For each timestep i, we denotes the state $s_i = [\mathbf{R}, \mathbf{p}, \mathbf{v}, \mathbf{b}]$ that consists of the rotation **R,** position $\mathbf{p}$, velocity estimate $\mathbf{v}$ and IMU bias. The pose ${}_B\boldsymbol{\xi} = (\mathbf{R}, \mathbf{p}) \in SE(3)$, $\mathbf{R} \in SO(3)$, $\mathbf{b} = [\mathbf{b}^g \ \mathbf{b}^a] \in \mathbb{R}^6$, $\mathbf{v}$, $\mathbf{b}^g$, $\mathbf{b}^a \in \mathbb{R}^3$, where $\mathbf{b}^g$, $\mathbf{b}^a$ are the gyroscope and accelerometer bias. IMU is typically composed of a three-axis accelerometer and a three-axis gyroscope that measures the acceleration and rotation rate of the IMU during motion. However, motion will lead to Gaussian white noise, therefore, the measurement model of IMU can be written as

$$\begin{aligned} {}_B\tilde{\boldsymbol{\omega}}_{WB}(t) &= {}_B\boldsymbol{\omega}_{WB}(t) + \mathbf{b}^g(t) + \boldsymbol{\eta}^g(t) \\ {}_W\tilde{\mathbf{a}}(t) &= \mathbf{R}^T_{WB}(t)({}_W\mathbf{a}(t) - {}_W\mathbf{g}) + \mathbf{b}^a(t) + \boldsymbol{\eta}^a(t) \end{aligned} \tag{10}$$

where W denotes the word coordinate, B denotes that the IMU coordinate, ${}_B\tilde{\omega}_{WB}(t)$ is the instantaneous angular velocity of B relative to W expressed in B coordinate, ${}_Wg$ is the gravity vector in W coordinate.

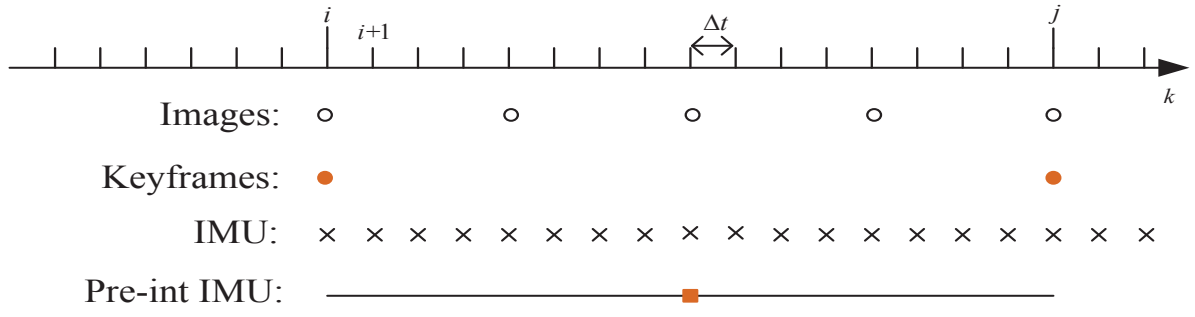


Figure 5: Different rates for IMU and camera.

To obtain IMU pose, the differential form of kinematic model in the word coordinate is

$$\begin{aligned} \dot{\mathbf{R}} &= \mathbf{R}\omega_{\times}, \\ \dot{\mathbf{v}} &= \mathbf{a}, \\ \dot{\mathbf{p}} &= \mathbf{v}. \end{aligned} \tag{11}$$

where $\omega_{\times}$ is the skew-symmetric metric.

Of course, it can also be written as a form of integral, and the discrete form of the integral model Eq. (11) at time $(t + \Delta t)$ is

$$\begin{aligned} \mathbf{R}_{WB}(t+\Delta t) &= \mathbf{R}_{WB}(t)\mathbf{Exp}({}_B\boldsymbol{\omega}_{WB}(t)\Delta t) \\ {}_W\mathbf{v}(t+\Delta t) &= {}_W\mathbf{v}(t) + {}_W\mathbf{a}(t)\Delta t \\ {}_W\mathbf{p}(t+\Delta t) &= {}_W\mathbf{p}(t) + {}_W\mathbf{v}(t)\Delta t + \frac{1}{2}{}_W\mathbf{a}(t)\Delta t^2 \end{aligned} \tag{12}$$

According to Eq. (10), we can obtain ${}_W\mathbf{a}$ and ${}_B\boldsymbol{\omega}_{WB}$. We drop the coordinate frame, and Eq. (12) is reweitten as follows

$$\begin{aligned} \mathbf{R}(t+\Delta t) &= \mathbf{R}(t)\mathbf{Exp}((\tilde{\boldsymbol{\omega}}(t) - \mathbf{b}^g(t) - \boldsymbol{\eta}^{gd}(t))\Delta t \\ \mathbf{v}(t+\Delta t) &= \mathbf{v}(t) + \mathbf{g}\Delta t + \mathbf{R}(t)(\tilde{\mathbf{a}}(t) - \mathbf{b}^a(t) - \boldsymbol{\eta}^{ad}(t))\Delta t \\ \mathbf{p}(t+\Delta t) &= \mathbf{p}(t) + \mathbf{v}(t)\Delta t + \frac{1}{2}\mathbf{g}\Delta t^2 + \frac{1}{2}\mathbf{R}(t)(\tilde{\mathbf{a}}(t) - \mathbf{b}^a(t) - \boldsymbol{\eta}^{ad}(t))\Delta t^2 \end{aligned} \tag{13}$$

where $\mathbf{Exp}(\cdot)$ denotes composite mapping corresponding to [25] (Eq. (6)), $\cdot^d$ denotes discrete, which is the Gaussian error of the measured value during Δt time period.

IMU measurements can reach a much higher frequency than the camera frame rate. Fig. 5 illustrates the difference between IMU measurement rates and camera frame rate. We do not add independent

residuals to each IMU measurements, instead, we integrate the measurements into a condensed IMU measurement between the image frames. If the pose or bias estimation changes during optimization, in order to avoid returning frequently, we follow the pre-integration method proposed in [25]. We assumed that the time of IMU measurements is Δt, and we integrate the IMU measurements at discrete time $k = i$ and $k = j$ (Fig. 5). The state value of j can be obtained from the state value integral at time i

$$\begin{aligned} \mathbf{R}_j &= \mathbf{R}_i \prod_{k=i}^{j-1} \mathbf{Exp}(\left(\tilde{\boldsymbol{\omega}}_k - b_k^g - \eta_k^{gd}\right) \Delta t) \\ \mathbf{v}_j &= \mathbf{v}_i + \mathbf{g}\Delta t_{ij} + \sum_{k=i}^{j-1} \mathbf{R}_k(\tilde{\mathbf{a}}_k - \mathbf{b}_k - \boldsymbol{\eta}_k^{ad})\Delta t \\ \mathbf{p}_j &= \mathbf{p}_i + \sum_{k=i}^{j-1}[\mathbf{v}_k\Delta t + \frac{1}{2}\mathbf{g}\Delta t^2 + \frac{1}{2}\mathbf{R}_k(\tilde{\mathbf{a}}_k - \mathbf{b}_k - \boldsymbol{\eta}_k^{ad})\Delta t^2] \end{aligned} \tag{14}$$

where $\Delta t_{ij} \doteq \sum_{k=i}^{j-1} \Delta t$.

Then the state at the next timestamp can be predicted. But in order to avoid to recompute $\mathbf{R}_k$ around each time integration, we use the relative motion increments to get the measurement model between two adjacent keyframes.

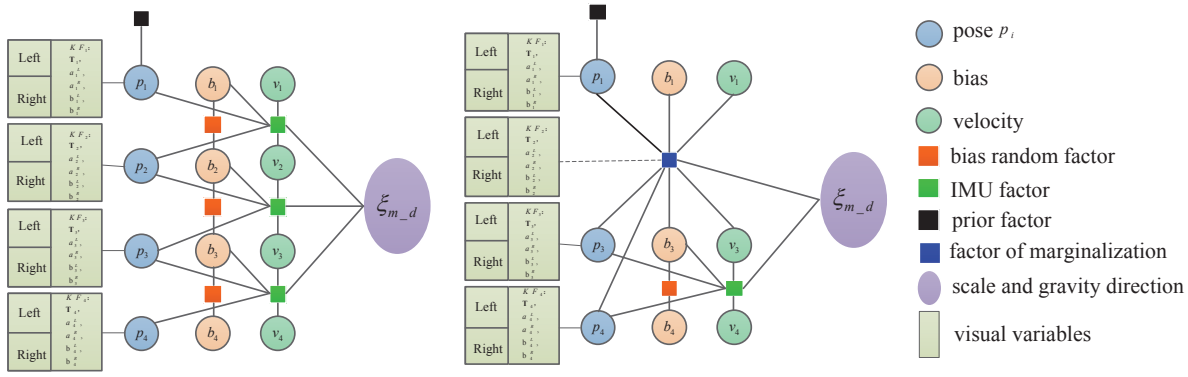
$$\begin{aligned} \Delta\mathbf{R}_{ij} &= \mathbf{R}_i^T\mathbf{R}_j = \prod_{k=i}^{j-1} \mathbf{Exp}(\left(\tilde{\boldsymbol{\omega}}_k - b_k^g - \eta_k^{gd}\right) \Delta t) \\ \Delta\mathbf{v}_{ij} &= \mathbf{R}_i^T(\mathbf{v}_j - \mathbf{v}_i - \mathbf{g}\Delta t_{ij}) = \sum_{k=i}^{j-1} \Delta\mathbf{R}_{ik}(\tilde{\mathbf{a}}_k - \mathbf{b}_k^a - \boldsymbol{\eta}_k^{ad})\Delta t \\ \Delta\mathbf{p}_{ij} &= \mathbf{R}_i^T(\mathbf{p}_j - \mathbf{p}_i - \mathbf{v}_i\Delta t_{ij} - \frac{1}{2}\mathbf{g}\Delta t_{ij}^2) = \sum_{k=i}^{j-1}[\Delta\mathbf{v}_{ik}\Delta t + \frac{1}{2}\Delta\mathbf{R}_{ik}(\tilde{\mathbf{a}}_k - \mathbf{b}_k^a - \boldsymbol{\eta}_k^{ad})\Delta t^2] \end{aligned} \tag{15}$$

The defined state quantities are independent of the state values $\mathbf{R}_i$, $\mathbf{v}_i$, $\mathbf{p}_i$, $\mathbf{R}_j$, $\mathbf{v}_j$, $\mathbf{p}_j$ at time i, j.

However, since the drift value b_k^g of the angular velocity measurement and the drift value b_k^a of the acceleration measurement at the intermediate time exit, for convenience, assume $b_k^g = b_i^g$, $b_k^a = b_i^a$, $k = i, i+1, \cdots, j-1$. Because bias also need to be estimated, this assumption reduces a large amount of the estimated state values. The measurement is related to the bias and noise, the bias is assumed to be known at time t_i. Then we can obtain

$$\begin{aligned} \Delta\mathbf{R}_{ij} &\doteq \Delta\tilde{\mathbf{R}}_{ij}\mathbf{Exp}(-\delta\phi_{ij}) \\ \Delta\mathbf{v}_{ij} &\doteq \Delta\tilde{\mathbf{v}}_{ij} - \delta\mathbf{v}_{ij} \\ \Delta\mathbf{p}_{ij} &\doteq \Delta\tilde{\mathbf{p}}_{ij} - \delta\mathbf{p}_{ij} \end{aligned} \tag{16}$$

where $\Delta\tilde{\mathbf{R}}_{ij} \doteq \prod_{k=i}^{j-1} \mathbf{Exp}((\tilde{\boldsymbol{\omega}}_k - b_k^g)\,\Delta t)$ is pre-intergrated rotation measurement, $\delta\phi_{ij}$, $\delta\mathbf{v}_{ij}$, $\delta\mathbf{p}_{ij}$ are the Gaussian errors corresponding to rotation, velocity and position respectively.



(a) Factor graph for the stereo visual- inertial optimization.

(b) Factor graph after keyframe 2 was marginalized.

Figure 6: Factor graph for the visual-inertial joint optimization before and after the marginalization of keyframes.

Thus IMU error can be obtained:

$$\begin{bmatrix} \delta\phi_{i,j} \\ \delta\mathbf{v}_{i,j} \\ \delta\mathbf{p}_{i,j} \end{bmatrix} = \begin{bmatrix} \Delta\tilde{\mathbf{R}}_{j-1,i}^{T}\delta\phi_{i,j-1} + \mathbf{J}_r^{j-1}\boldsymbol{\eta}_{j-1}^{ad}\Delta t \\ \delta\mathbf{v}_{i,j-1} - \Delta\tilde{\mathbf{R}}_{i.j-1}(\tilde{\mathbf{a}}_{i,j-1} - \mathbf{b}_i^a)\delta\phi_{i,j-1}\Delta t + \Delta\tilde{\mathbf{R}}_{i.j-1}\boldsymbol{\eta}_{j-1}^{ad}\Delta t \\ \delta\mathbf{p}_{i,j-1} + \delta\mathbf{v}_{i,j-1}\Delta t - \frac{1}{2}\Delta\tilde{\mathbf{R}}_{i.j-1}(\tilde{\mathbf{a}}_{i,j-1} - \mathbf{b}_i^a)\delta\phi_{i,j-1}\Delta t^2 + \frac{1}{2}\Delta\tilde{\mathbf{R}}_{i.j-1}\boldsymbol{\eta}_{j-1}^{ad}\Delta t^2 \end{bmatrix} \tag{17}$$

The bias from frame i to j will be updated by calculating the Jacobian matrix of the bias in Eq. (15) to Eq. (16), and the bias error terms are:

$$\begin{aligned} \delta\mathbf{b}_{i,j}^g &= \mathbf{b}_{i,j}^g - \mathbf{b}_{i,j-1}^g \\ \delta\mathbf{b}_{i,j}^a &= \mathbf{b}_{i,j}^a - \mathbf{b}_{i,j-1}^a \end{aligned} \tag{18}$$

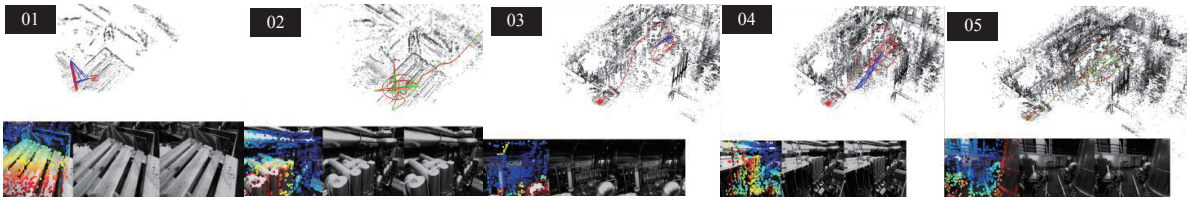
The error energy function is

$$E_{IMU}(s_i, s_j) = \begin{bmatrix} \delta\phi_{i,j} \\ \delta\mathbf{v}_{i,j} \\ \delta\mathbf{p}_{i,j} \\ \delta\mathbf{b}_{i,j}^g \\ \delta\mathbf{b}_{i,j}^a \end{bmatrix} := (s_j \boxminus \hat{s}_j)^T \sum\nolimits_{i,j-1}^{-1} (s_j \boxminus \hat{s}_j) \tag{19}$$

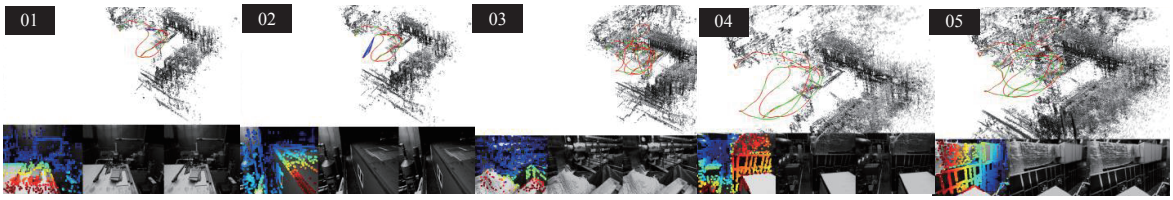
where $\sum_{i,j-1}^{-1}$ is the associated covariance matrix, the $\boxminus$ obeys $\xi_j \boxplus (\hat{\xi}_j)^{-1}$.

3.4. Window optimization

We optimize the poses, IMU-biases and velocities of the fixed number of the keyframes. Fig. 6(a) shows a factor graph. Note that there should be several visual factors between the two keyframes. Each



(a) EuRoC MH_01 Dataset



(b) EuRoC MH_03 Dataset

Figure 7: EuRoC Dataset. My method (stereo_VIDSO Algorithom) generates a consistent global map, the semi-dense depth map of using direct alignment and probabilistic instead of feature points method.

IMU factor connecting two successive keyframes based on pre-integration is described in Section 3.3. Because the errors of pre-integration increases with time, the proposed method can ensure a time lower than 0.5s, and the marginalization process relationship is observed. In this paper, the result appears in the stereo DSO frame rather than in the metric frame.

3.4.1. Nonlinear optimization formulation

We define a state vector

$$s = [c^T, \xi_{i-1}^T, a_i, b_i, s_i^T]^T \tag{20}$$

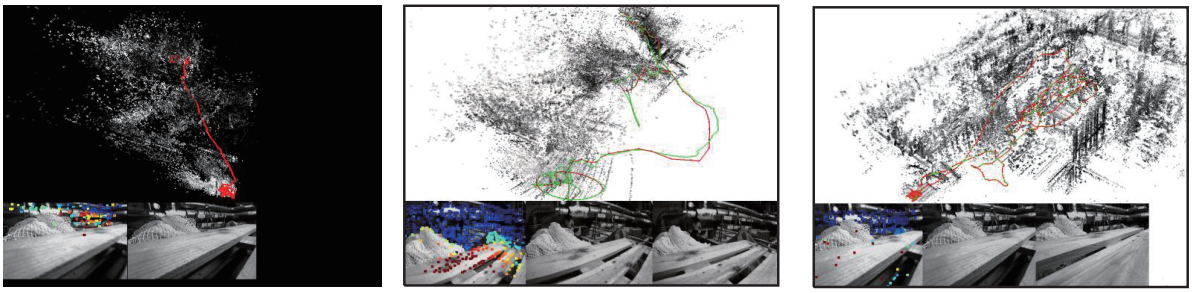
where c contains intrinsic parameters of the camera, and ξ_{i-1} is the prior camera pose, a_i, b_i are the affine illumination parameters, and $s_i = [{}_B\boldsymbol{\xi}, \mathbf{v}_i, \mathbf{b}_i]$ is the current IMU state.

We perform nonlinear optimization using the Gauss-Newton system $\mathbf{H}\delta = \mathbf{b}$. The error in Eq. (1) can be written as

$$\begin{aligned} E &= \frac{\lambda}{2} r^T \mathbf{W} r \\ &= \frac{\lambda}{2} [r_I^T \; r_{IMU}^T] \begin{bmatrix} \mathbf{W}_I & 0 \\ 0 & \mathbf{W}_{IMU} \end{bmatrix} \begin{bmatrix} r_I \\ r_{IMU} \end{bmatrix} \end{aligned} \tag{21}$$

where $\mathbf{W}$ is a diagonal weight matrix.

We define $s \boxplus s'$ to obey the operation $\xi \boxplus \xi'$ for the Lie algebra components.



(a) DSO Algorithom (b) Stereo DSO Algorithom (c) Stereo VI DSO Algorithm

Figure 8: Qualitative results on EuRoC dataset.

$\mathbf{J}$ is defined using the stacked residual vector r as

$$\mathbf{J} = \frac{dr(s \boxplus \delta s)}{d\delta s}|_{\delta s=0},\ \mathbf{H} = \mathbf{J}^T\mathbf{W}\mathbf{J},\ \mathbf{b} = -\mathbf{J}^T\mathbf{W}\mathbf{r} \tag{22}$$

where $\mathbf{b}$ is the Jacobian and $\mathbf{H}$ is the Hessian of E (Eq. (21)). Then we use $\delta = \mathbf{H}^{-1}\mathbf{b}$ to update.

In fact, the camera photometric error E_{stereo} and IMU energy error E_{IMU} do not have common bias and residuals. Therefore, we divide $\mathbf{H}$ and $\mathbf{b}$ into two independent terms

$$\mathbf{H} = \mathbf{H}_{stereo} + \mathbf{H}_{IMU}, \mathbf{b} = \mathbf{b}_{stereo} + \mathbf{b}_{IMU}. \tag{23}$$

Since the current relative pose estimation originates from inertia residuals, it is necessary to use the IMU relative pose in the metric framework. The IMU residuals cause

$$\mathbf{H}'_{IMU} = \mathbf{J}'^T_{IMU}\mathbf{W}_{IMU}J'_{IMU}, \mathbf{b}'_{IMU} = -\mathbf{J}'^T_{IMU}\mathbf{W}_{IMU}\mathbf{r}. \tag{24}$$

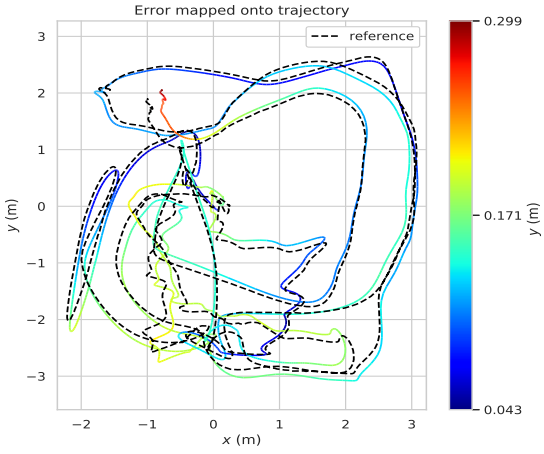
However, in the joint optimization frame, we need to obtain $\mathbf{H}_{IMU}$ and $\mathbf{b}_{IMU}$ based on the state definition in Eq. (23). According to the difference of the two states in the pose representation, we can obtain $\mathbf{J}_{rel}$

$$\mathbf{H}_{IMU} = \mathbf{J}^T_{rel} \cdot \mathbf{H}'_{IMU} \cdot \mathbf{J}_{rel}\ , \mathbf{b}_{IMU} = \mathbf{J}^T_{rel} \cdot \mathbf{b}'_{IMU} \tag{25}$$

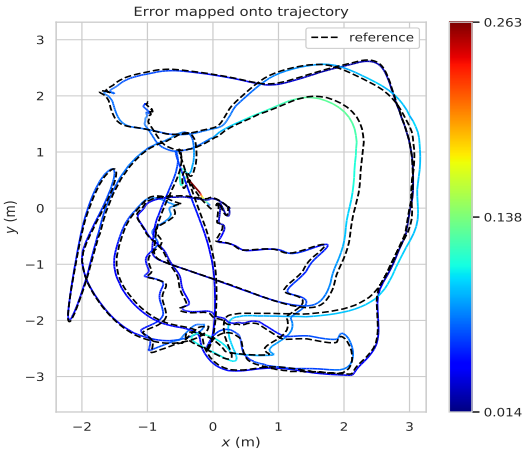
The computation of $\mathbf{J}_{rel}$ refers to [19].

3.4.2. Marginalization

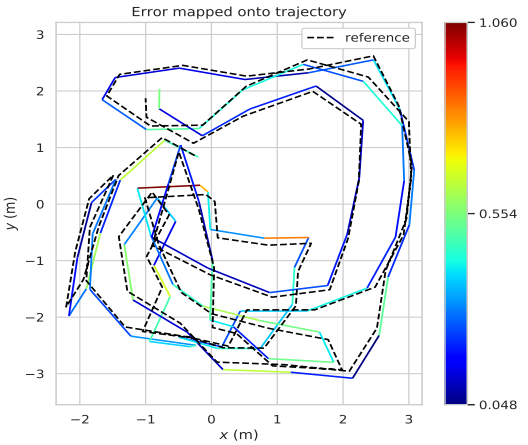
Fig. 6(b) shows the procedure of marginalization. In order to achieve a Gauss-Newton update, we perform the marginalization for older keyframes. This process means that all variables corresponding to the current frame (including pose and velocity) are marginalized out using the Schur complement.



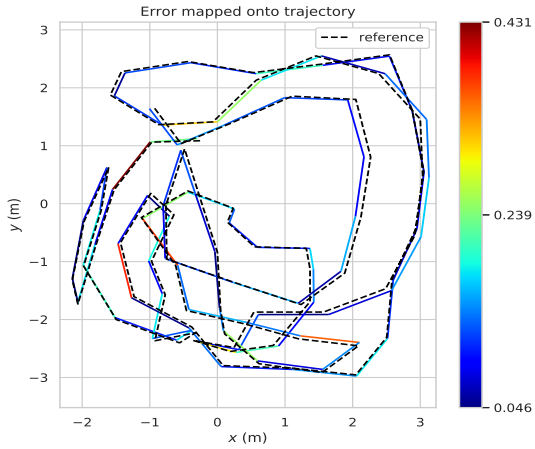
(a) APE error of mono_VIDSO



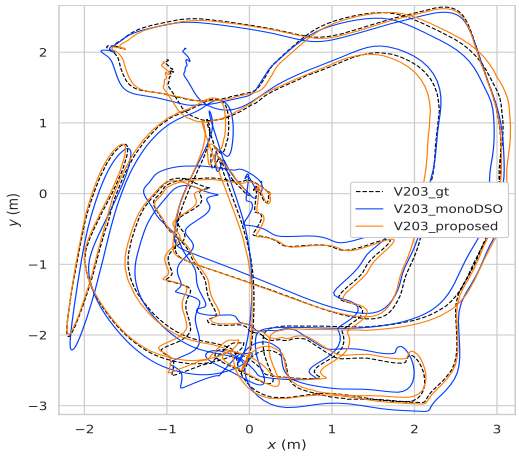
(b) APE error of proposed



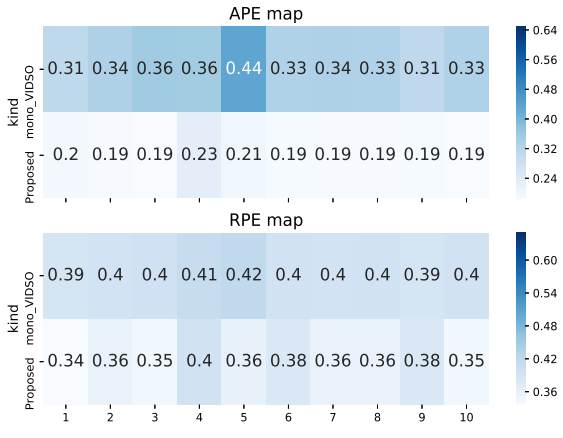
(c) RPE error of mono_VIDSO



(d) RPE error of proposed



(e) Trajectory



(f) APE and RPE run 10 times

Figure 9: The qualitative compare results on EuRoC V203 dataset. In the legend, the V203_MonoDSO is the Mono_VIDSO method.

In visual factor marginalization, the remaining items affecting the system sparsity are deleted, and all points are first marginalized in the keyframe before the keyframe itself is marginalized [4]. Then, the keyframe is marginalized and moved out of the active windows. Because the factors caused by marginalization require that the linearization points of all connected variables remain fixed, we apply the method in [4] approximately to further linearize the energy around the linearization point. In order to maintain system consistency, it is important that the Jacobian evaluates the same value for the variables associated with the marginalization factor. Otherwise the zero space is eliminated. Thus, we adopt "First Estimates Jacobians." If we use s_α to denote the state variables that we want to keep in the optimization and denote s_β that we want to marginalized out, the Gaussian-Newton system can be write as follows

$$\begin{bmatrix} \mathbf{H}_{\alpha\alpha} & \mathbf{H}_{\alpha\beta} \\ \mathbf{H}_{\beta\alpha} & \mathbf{H}_{\beta\beta} \end{bmatrix} \begin{bmatrix} \delta s_\alpha \\ \delta s_\beta \end{bmatrix} = \begin{bmatrix} \mathbf{b}_\alpha \\ \mathbf{b}_\beta \end{bmatrix} \tag{26}$$

Multiply the second line by $\mathbf{H}_{\alpha\beta}\mathbf{H}_{\beta\beta}^{-1}$ and subtract it from the first element, then the Eq. (26) becomes

$$\underbrace{(\mathbf{H}_{\alpha\alpha} - \mathbf{H}_{\alpha\beta}\mathbf{H}_{\beta\beta}^{-1}\mathbf{H}_{\alpha\beta})}_{\hat{\mathbf{H}}_{\alpha\alpha}}\delta s_\alpha = \underbrace{\mathbf{b}_\alpha - \mathbf{H}_{\alpha\beta}\mathbf{H}_{\beta\beta}^{-1}\mathbf{b}_\beta}_{\hat{\mathbf{b}}_\alpha} \tag{27}$$

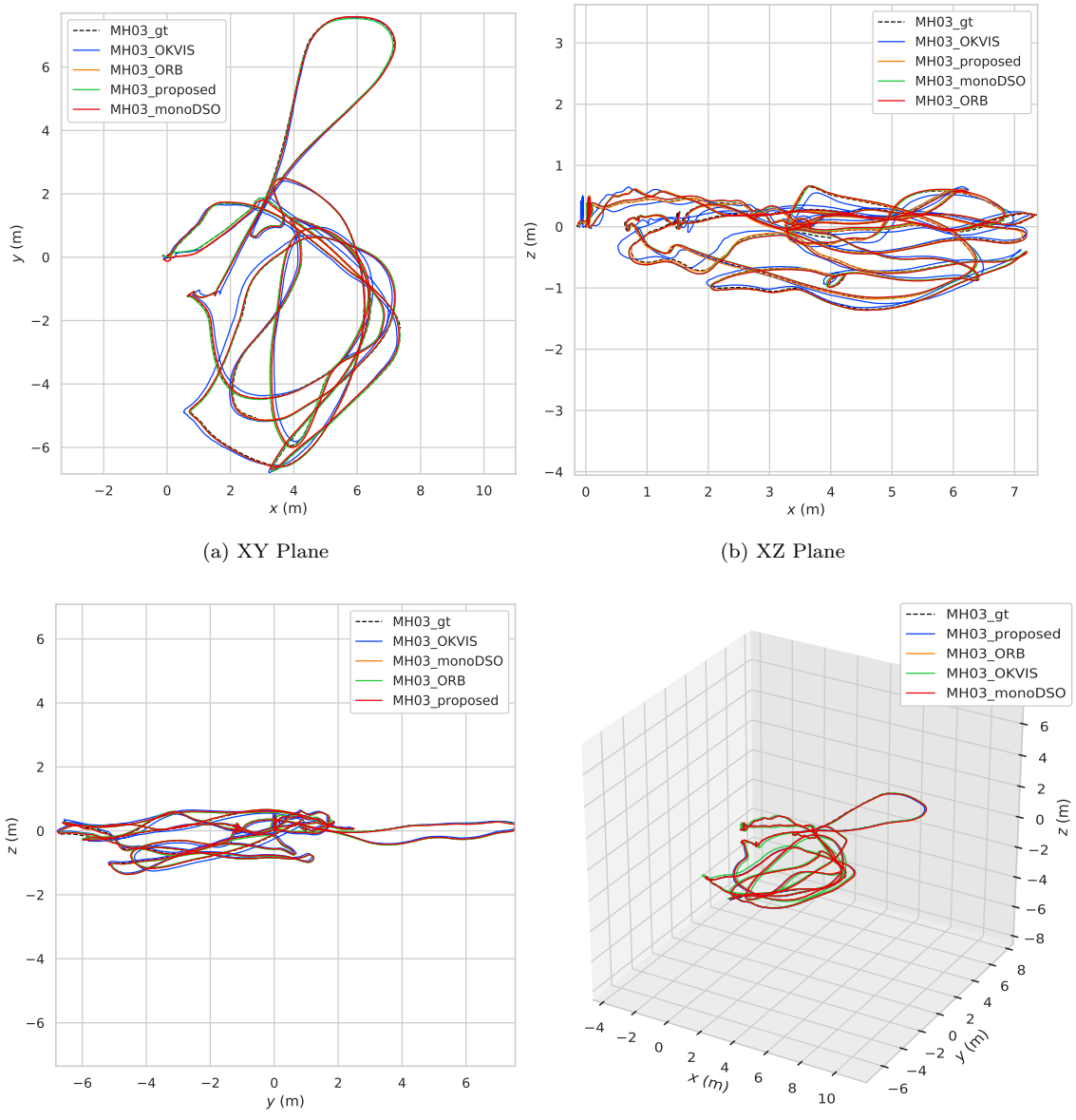
Then the states s_β marginalized out but the information of s_β will be preserved and utilized.

4. Results

We evaluated our method using the public EuRoC dataset (see Fig. 7). The images were provided by the required calibration parameters and groundtruth based on motion capture. The dataset contained two calibrated stereo video sequences corresponding to IMU measurements, and was recorded using a Skybotix VI sensor. We compared it with DSO, stereo DSO and other visual-inertial system, and evaluated the effect of the proposed method and parameters selection.

4.1. Qualitative comparison on large trajectories

Since the EuRoC datasets sequences set moves faster, there is large jitter and low partial visibility, which is difficult for visual-only SLAM. The proposed method in this paper improves the above problem on the basis of DSO. Fig. 8 shows the comparison results of 3D reconstruction effects for the proposed stereo_VIDSO method, visual-only DSO and stereo DSO. The results show that the proposed method is more effective than the visual-only method. In Fig. 8(a), Fig. 8(b), the 3D reconstruction of stereo



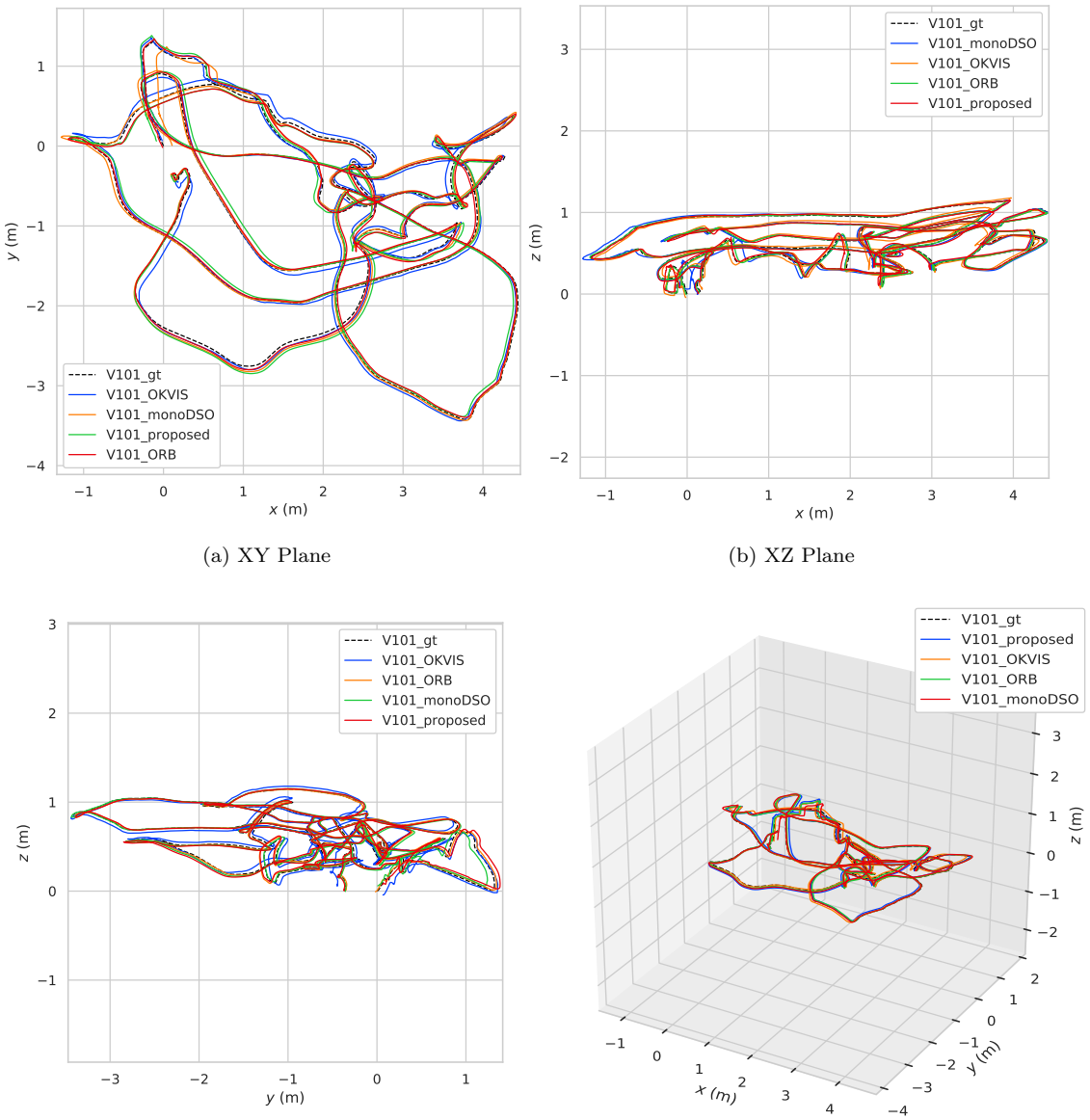
(a) XY Plane

(b) XZ Plane

(c) YZ Plane

(d) 3D Plane

Figure 10: The estimated trajectory results on EuRoC MH03 dataset. In the legend, MH03_ORB is the stereo_VIORB method, the MH03_MonoDSO is the Mono_VIDSO method.



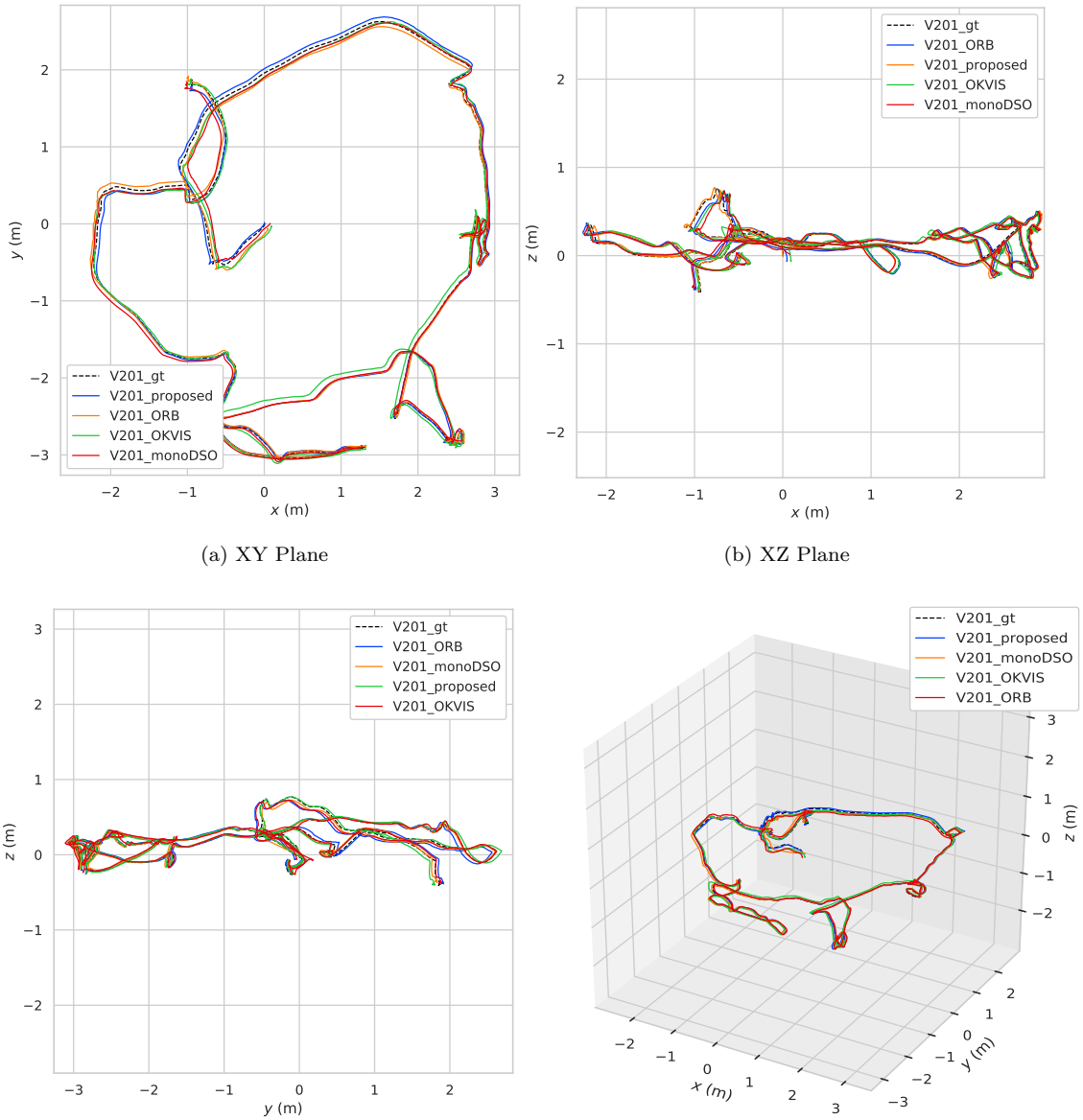
(a) XY Plane

(b) XZ Plane

(c) YZ Plane

(d) 3D Plane

Figure 11: The estimated trajectory results on EuRoC V101 dataset. In the legend, V101_ORB is the stereo_VIORB method, the V101_MonoDSO is the Mono_VIDSO method.



(a) XY Plane

(b) XZ Plane

(c) YZ Plane

(d) 3D Plane

Figure 12: The estimated trajectory results on EuRoC V201 dataset. In the legend, V201_ORB is the stereo_VIORB method, the V201_MonoDSO is the Mono_VIDSO method.

Table 1: Accuracy of the RMSE on the EuRoC dataset.

Sequence	OKVIS	Stereo_VIORB	Mono_VIDSO	Proposed
MH_01_easy	0.190942	0.041853	0.034971	**0.019273**
MH_02_easy	0.109778	0.033052	0.041853	**0.029431**
MH_03_medium	0.144528	0.054798	0.059897	**0.032990**
$MH\_04\_difficult$	0.181759	0.126432	0.1327143	**0.102714**
$MH\_05\_difficult$	0.079129	**0.049930**	0.082149	0.066954
$V1\_01\_easy$	0.059106	0.084286	0.063199	**0.055838**
$V1\_02\_medium$	**0.039165**	0.060481	0.019187	0.120850
$V1\_03\_difficult$	0.125028	**0.067898**	0.300904	0.169414
$V2\_01\_easy$	0.061776	0.050130	0.075491	**0.041523**
$V2\_02\_medium$	0.094641	0.078541	0147733	**0.046415**
$V2\_03\_difficult$	1.921737	0.357904	0.166868	**0.155817**

DSO method outperforms the 3D reconstruction of DSO. The scene information in Fig. 8(c) is more clearly than that of in Fig. 8a and Fig. 8(b). The stereo vision SLAM system recovers the scale from static stereo matching, solves the problem of scale uncertainty in the monocular SLAM system, and improves the accuracy and robustness. In order to achieve the comparison, we also modified the program to achieve monocular visual-inertial DSO (mono_VIDSO). Fig. 9 shows the comparison between the proposed method and the mono_VIDSO [19] method. We use the absolute pose error (APE) and relative pose error (RPE) metrics to evaluate the competing algorithms. APE is often used to measure estimation error along a trajectory. The estimated value and reference value are directly compared if pose correspondence is given. Then statistics for the whole trajectory are calculated, and they are used to measure the global performance of a trajectory. The lower the APE is, the better the performance. The comparison results in terms of APE are shown in Fig. 9(a) and Fig. 9(b). RPE reflects the local localization accuracy. We calculate the RPE every one meter. In Fig. 9(c) and Fig. 9 (d), error magnitude is coded in pseudocolor. Fig. 9(e) shows the trajectories estimated by mono_VIDSO and the proposed method with respect to the reference trajectory, and we can obviously see that the trajectory of the proposed method is closer to the groundtruth. In order to directly compare the accuracy of the proposed method with the monocular method in terms of scale recovery, we draw the APE and RPE thermodynamic comparison chart. The smaller the color difference is, the more similar the value. The

Table 2: Accuracy of the APE on the EuRoC dataset.

Sequence	OKVIS	Stereo_VIORB	Mono_VIDSO	Proposed
MH_01_easy	0.164533	0.039492	0.031964	**0.016544**
MH_02_easy	0.325944	0.071015	0.227868	**0.066134**
MH_03_medium	0.135089	0.047799	0.053473	**0.029635**
$MH\_04\_difficult$	0.177335	0.076432	0.089257	**0.072714**
$MH\_05\_difficult$	0.072911	**0.059745**	0.072949	0.060715
$V1\_01\_easy$	0.055121	0.052416	0.051721	**0.051811**
$V1\_02\_medium$	**0.034487**	0.060481	0.055843	0.055441
$V1\_03\_difficult$	0.212638	**0.133869**	0.212638	0.136592
$V2\_01\_easy$	0.055524	0.054945	0.055301	**0.036858**
$V2\_02\_medium$	0.088461	0.059489	0.126172	**0.039465**
$V2\_03\_difficult$	1.741872	0.152617	0.155504	**0.139130**

colder the color is, the smaller the scale error.

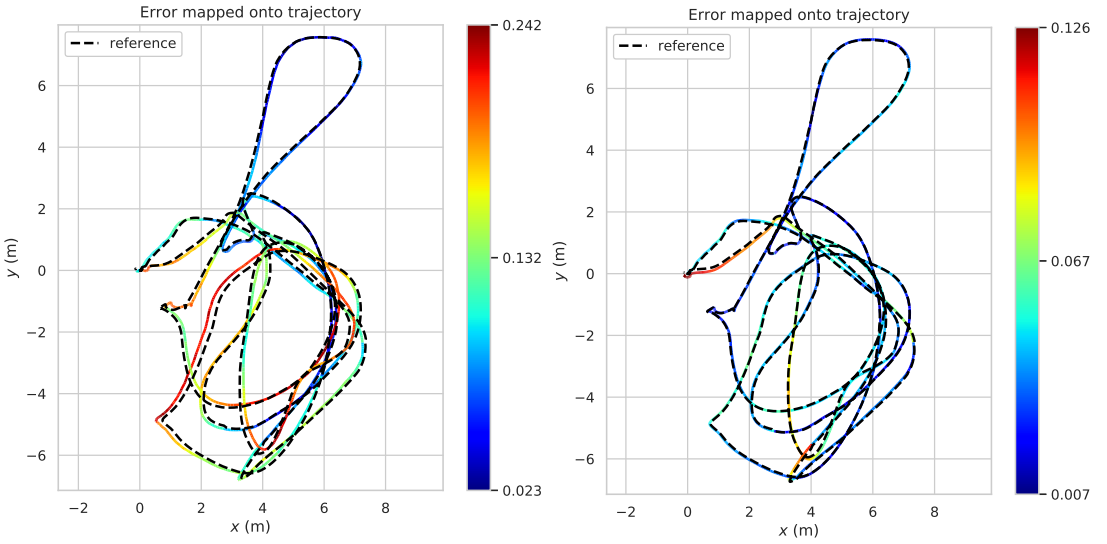
4.2. Long term drift evaluation

In this section, we compare the performance of our proposed method with state-of-the-art SLAM algorithms that use IMU, such as stereo_VIORB SLAM [20] and OKVIS [15]. We study monocular inertial DSO (named mono_VIDSO in this paper) [19] and compare it with our proposed method. Fig. 10-12 show that the tightly integrated and direct VIO method proposed in this paper outperforms several other methods in the three indoor datasets. The video datasets MH_03, V1_01 and V2_01 exist obvious dithering at the beginning of the trajectories, and IMU is able to respond to this drift. In these figures, the groundtruth of EuRoC dataset is obtained from GPS and other sensors. They show the superiority of our proposed algorithm. The trajectory estimated by OKVIS and stereo_VIORB is close to the groundtruth trajectory, which demonstrates that the pose estimated by the stereo camera is better than that by monocular camera. The trajectory estimated by our method is the closest to the groundtruth, and this demonstrates that the proposed method has better robustness than mono_VIDSO, OKVIS and stereo_VIORB.

Fig. 13 shows the results of the proposed method, stereo_VIORB, OKVIS and mono_VIDSO compared with groundtruth. The maximum trajectory error of OKVIS is 0.242, the minimum trajectory

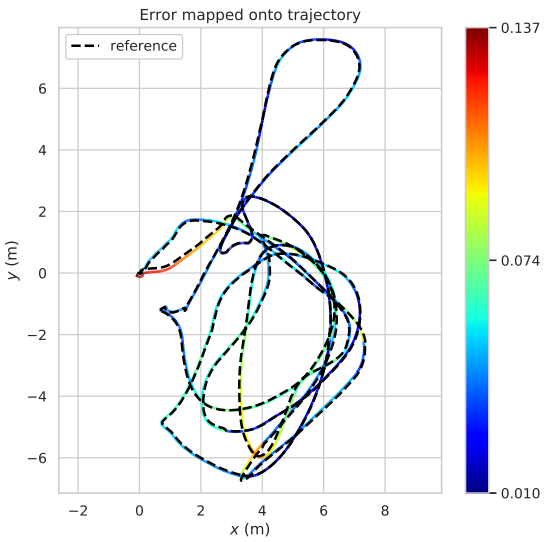
error is 0.023, and the mean error is 0.132. Fig. 13 (b), Fig. 13(c) and Fig. 13(d) are the trajectory errors of the stereo_VIORB, OKVIS, mono_VIDSO and proposed method, respectively.

Fig. 14 shows the APE values of stereo_VIORB, OKVIS, mono_VIDSO and the proposed method compared with groundtruth. Fig. 14(a) compares the four methods in terms of APE values and Fig. 14(b) is the box chart of APE. Fig. 14(c) compares the four methods in terms of the error statistics of standard deviation (std), root mean square error (rmse), minimum, median, mean and maximum. Based on these results, one can conclude that the performance of the proposed method is better than stereo_VIORB, OKVIS, mono_VIDSO. In order to obtain an accurate evaluation, we run each sequence of the MH03 dataset 10 times using our method. We compare the proposed method with stereo_VIORB, OKVIS, mono_VIDSO, and the results prove the robustness of our method. Fig. 15(a) is the root mean square error (RMSE) and Fig. 15(b) is the RPE. The RMSE is the square root that is the ratio of the sum of the error square for the observed value, the true value and the observation times. It is used to measure the deviation between the observed value and the true value. The smaller the RMSE value is, the smaller the deviation. Fig. 15(a) demonstrates that the RMSE value of the proposed method is the smallest at around 0.03. Both the mono_VIDSO and stereo_VIORB have RMSE of around 0.06. The OKVIS's RMSE error is much larger at 0.16. We attribute the superior performance of our method to the introduction of IMU that can provide a reliable scale estimate. Fig. 15(b) shows that the error of the proposed method is the smallest among all competing methods. To further prove the superiority of our algorithm, we run experiments on all the EuRoC dataset sequences. Table 1 shows the RMSE values compared with stereo_VIORB SLAM, OKVIS, mono_VIDSO. Table 1 demonstrates that the proposed method can perform well in the most of the dataset sequences. However, in the MH_05 and V1_03, stereo_VIORB method performs better than our method, and OKVIS method performs better in the V1_02. The main reason is that stereo_VIORB and OKVIS use loop-closures. Table 2 shows the comparison in terms of RPE. The proposed method performs well in the most of the dataset sequences. However, in the MH_03, V1_01 and V2_01, stereo VIORB method performs better than the proposed method, and OKVIS method performs better in the V1_02. The main reason is that our method selects 2000 map point to optimize in each frame, and leads to local localization bad in a scene without significant gradient changes. Finally, to tested the system thoroughly, we also run on KITTI dataset sequence. Our method can real-time tracking of stereo frames and build a 3D map. Fig. 16 shows the result of our method on KITTI dataset.

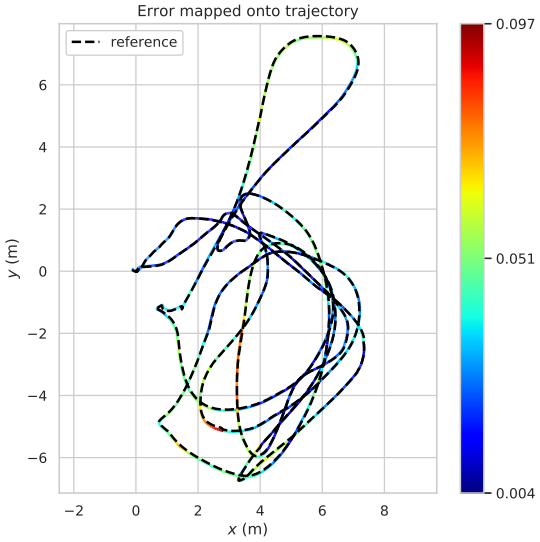


(a) OKVIS error.

(b) Stereo_VIORB error

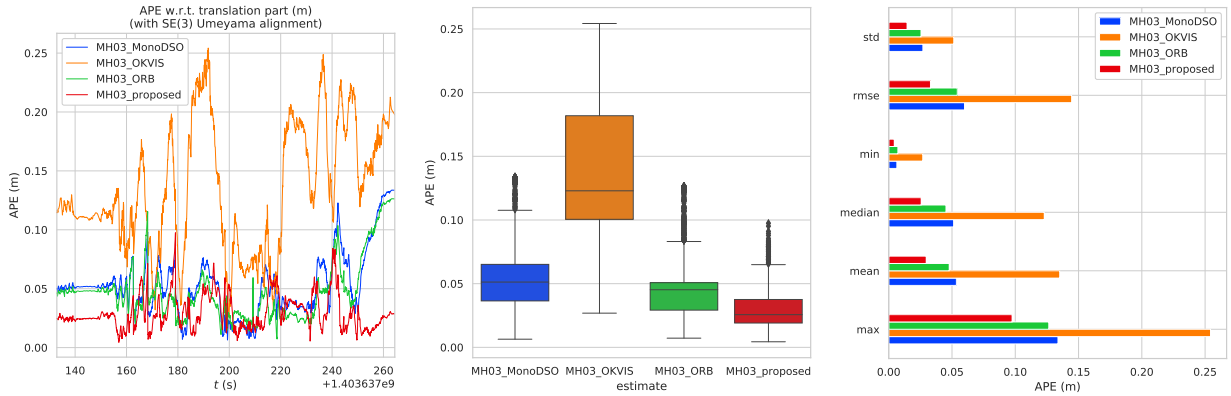


(c) Mono_VIDSO error



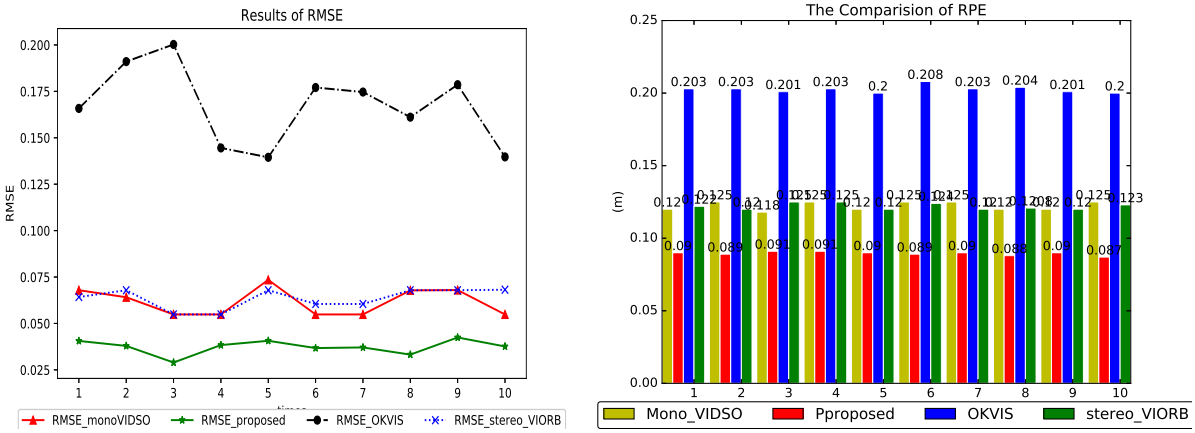
(d) Proposed method error

Figure 13: The comparison of trajectory errors between these four methods and groundtruth on EuRoC MH_03 dataset.



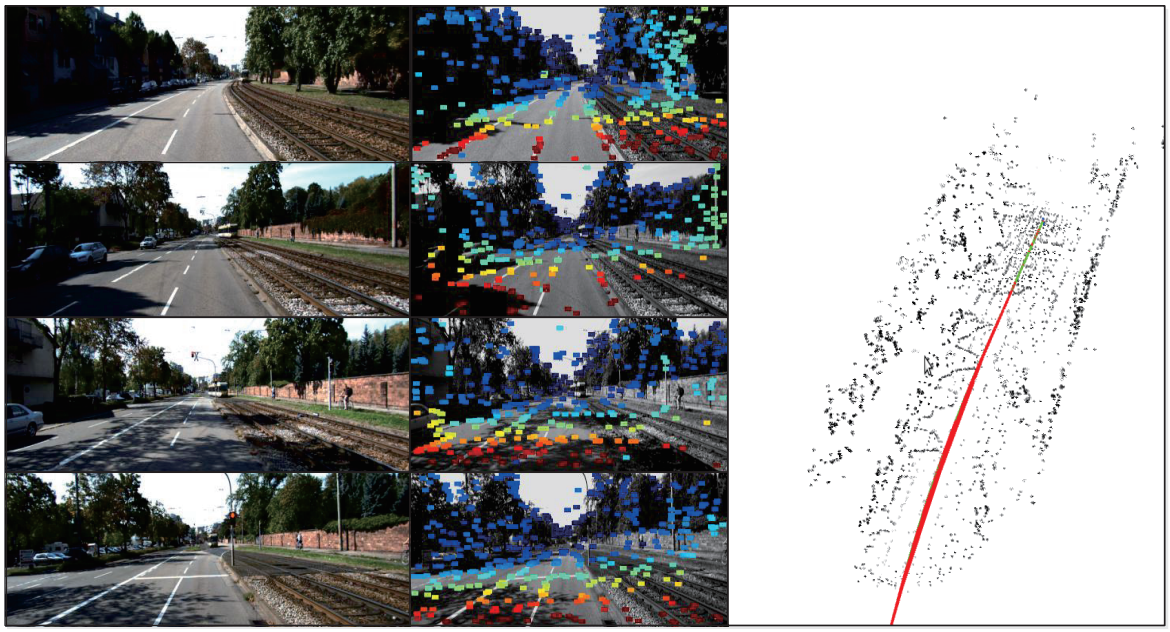
(a) The comparison chart of APE. (b) The box shape chart of APE. (c) The chart of different parameters.

Figure 14: The comparison results of APE on EuRoC MH_03 dataset.



(a) The comparison chart of RMSE. (b) The box shape chart of RPE.

Figure 15: The errors results in EuRoC MH_03 dataset.



(a) Original image (b) Depth image (c) Semi-dense map

Figure 16: The result of KITTI (2011.9.26_drive_0001_sync) dataset.

5. Conclusion

We proposed a novel direct sparse stereo vision-inertial odometry method. The fusion of an IMU and stereo vision can compensate for their individual disadvantages. Stereo vision allows the system to compensate for long-term IMU bias drift, while short-term IMU constraints can overcome non-convexity in the photometric tracking formula, and allow for large inter-frame motion or interval tracking without visual information. The proposed method created a semi-dense map with good observability, and accurate 3D reconstruction of the environment. In this paper, we focused on the front-end design, according to the analysis of the experimental results, the error will accumulate during the SLAM process. In order to solve this problem, we will work on the SLAM back-end that can improve the robustness of the optimization with loop closure detection in the future work. In this paper, we have focused on indoor localization and mapping in the experimental evaluation of our method using EuRoC datasets and outdoor. However, we will consider closure detection in the future work.

6. Acknowlegements

The work is supported by the national Natural Science Foundation of China (Project No. 61673125, 61773333), China Scholarship Council (CSC, Project No.201908130016).

References

[1] R. Urtasun, P. Lenz, A. Geiger, Are we ready for autonomous driving? The KITTI vision benchmark suite, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3354–3361.

[2] V. Usenko, J. Engel, J. Stckler, D. Cremers, Direct visual-inertial odometry with stereo cameras, in: IEEE International Conference on Robotics and Automation, 2016, pp. 1885–1892.

[3] S. L. Bowman, N. Atanasov, K. Daniilidis, G. J. Pappas, Probabilistic data association for semantic SLAM, in: IEEE International Conference on Robotics and Automation, 2017, pp. 1722–1729.

[4] J. Engel, V. Koltun, D. Cremers, Direct Sparse Odometry, IEEE Transactions on Pattern Analysis and Machine Intelligence PP (99) (2016) 1–1.

[5] A. I. Comport, E. Malis, P. Rives, Accurate Quadrifocal Tracking for Robust 3D Visual Odometry, in: IEEE International Conference on Robotics and Automation, 2007, pp. 40–45.

[6] C. Kerl, J. Sturm, D. Cremers, Robust odometry estimation for RGB-D cameras, in: IEEE International Conference on Robotics and Automation, 2013, pp. 3748–3754.

[7] R. A. Newcombe, S. J. Lovegrove, A. J. Davison, DTAM: Dense tracking and mapping in real-time, in: International Conference on Computer Vision, 2010, pp. 2320–2327.

[8] C. Forster, M. Pizzoli, D. Scaramuzza, SVO: Fast semi-direct monocular visual odometry, in: IEEE International Conference on Robotics and Automation, 2014, pp. 15–22.

[9] J. Engel, T. Schps, D. Cremers, LSD-SLAM: Large-Scale Direct Monocular SLAM, Springer International Publishing, 2014.

[10] N. Mahmoud, I. Cirauqui, A. Hostettler, C. Doignon, L. Soler, J. Marescaux, J. M. M. Montiel, ORBSLAM-Based Endoscope Tracking and 3D Reconstruction.

[11] Mur-Artal, Raul and Montiel, J. M. M. and Tardos, Juan D., ORB-SLAM: a Versatile and Accurate Monocular SLAM System, IEEE Transactions on Robotics 31 (5) (2017) 1147–1163.

[12] G. Klein, D. Murray, Parallel Tracking and Mapping for Small AR Workspaces, in: IEEE and ACM International Symposium on Mixed and Augmented Reality, 2008, pp. 1–10.

[13] J. Engel, J. Sturm, D. Cremers, Camera-based navigation of a low-cost quadrocopter, in: Ieee/rsj International Conference on Intelligent Robots and Systems.

[14] L. Meier, P. Tanskanen, F. Fraundorfer, M. Pollefeys, The Pixhawk Open-Source Computer Vision Framework for Mavs, ISPRS - International Archives of the Photogrammetry XXXVIII-1/C22 (2012) 13–18.

[15] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, P. Furgale, Keyframe-based visual-inertial odometry using nonlinear optimization, International Journal of Robotics Research 34 (3) (2014) 314–334.

[16] T. Qin, P. Li, S. Shen, VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator, IEEE Transactions on Robotics PP (99) (2017) 1–17.

[17] R. Mur-Artal, J. D. Tards, Visual-Inertial Monocular SLAM With Map Reuse, IEEE Robotics and Automation Letters 2 (2) (2016) 796–803.

[18] J. Engel, V. Usenko, D. Cremers, A Photometrically Calibrated Benchmark For Monocular Visual Odometry arXiv:1607.02555[cs.CV].

[19] L. V. Stumberg, V. Usenko, D. Cremers, Direct Sparse Visual-Inertial Odometry using Dynamic Marginalization, arXiv:1804.05625[cs.CV].

[20] R. Mur-Artal, J. D. Tardos, ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras, IEEE Transactions on Robotics 33 (5) (2017) 1255–1262.

[21] R. Wang, M. Schworer, D. Cremers, Stereo DSO: Large-Scale Direct Sparse Visual Odometry with Stereo Cameras (2017) 3923–3931,doi:10.1109/ICCV.2017.421.

[22] C. Forster, L. Carlone, F. Dellaert, D. Scaramuzza, On-Manifold Preintegration for Real-Time Visual–Inertial Odometry, IEEE Transactions on Robotics 33 (1) (2015) 1–21.

[23] J. Engel, J. Stckler, D. Cremers, Large-scale direct SLAM with stereo cameras, in: IEEE/rsj International Conference on Intelligent Robots and Systems, 2015, pp. 1935–1942.

[24] J. Engel, J. Sturm, D. Cremers, Semi-dense Visual Odometry for a Monocular Camera, in: IEEE International Conference on Computer Vision, 2013, pp. 1449–1456.

[25] C. Forster, L. Carlone, F. Dellaert, D. Scaramuzza, On-Manifold Preintegration for Real-Time Visual–Inertial Odometry, IEEE Transactions on Robotics 33 (1) (2017) 1–21.